

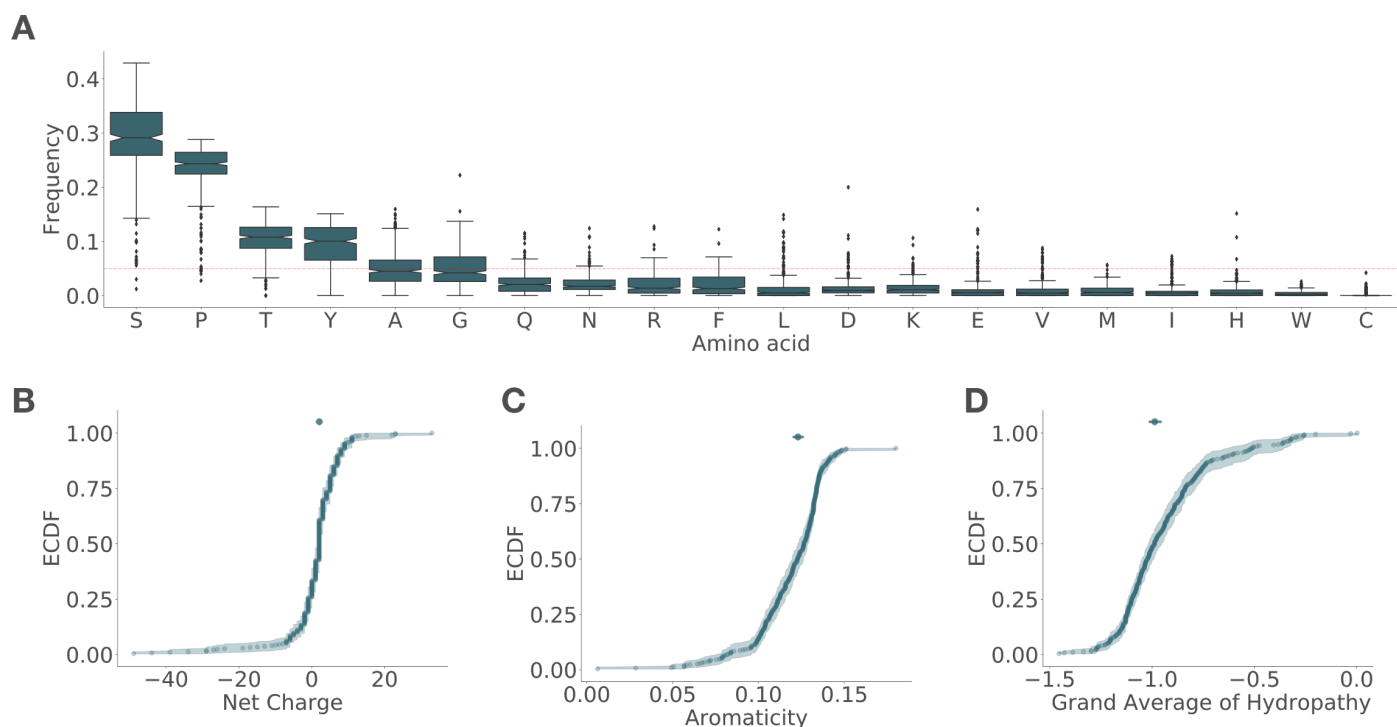
**Molecular Cell, Volume 79**

**Supplemental Information**

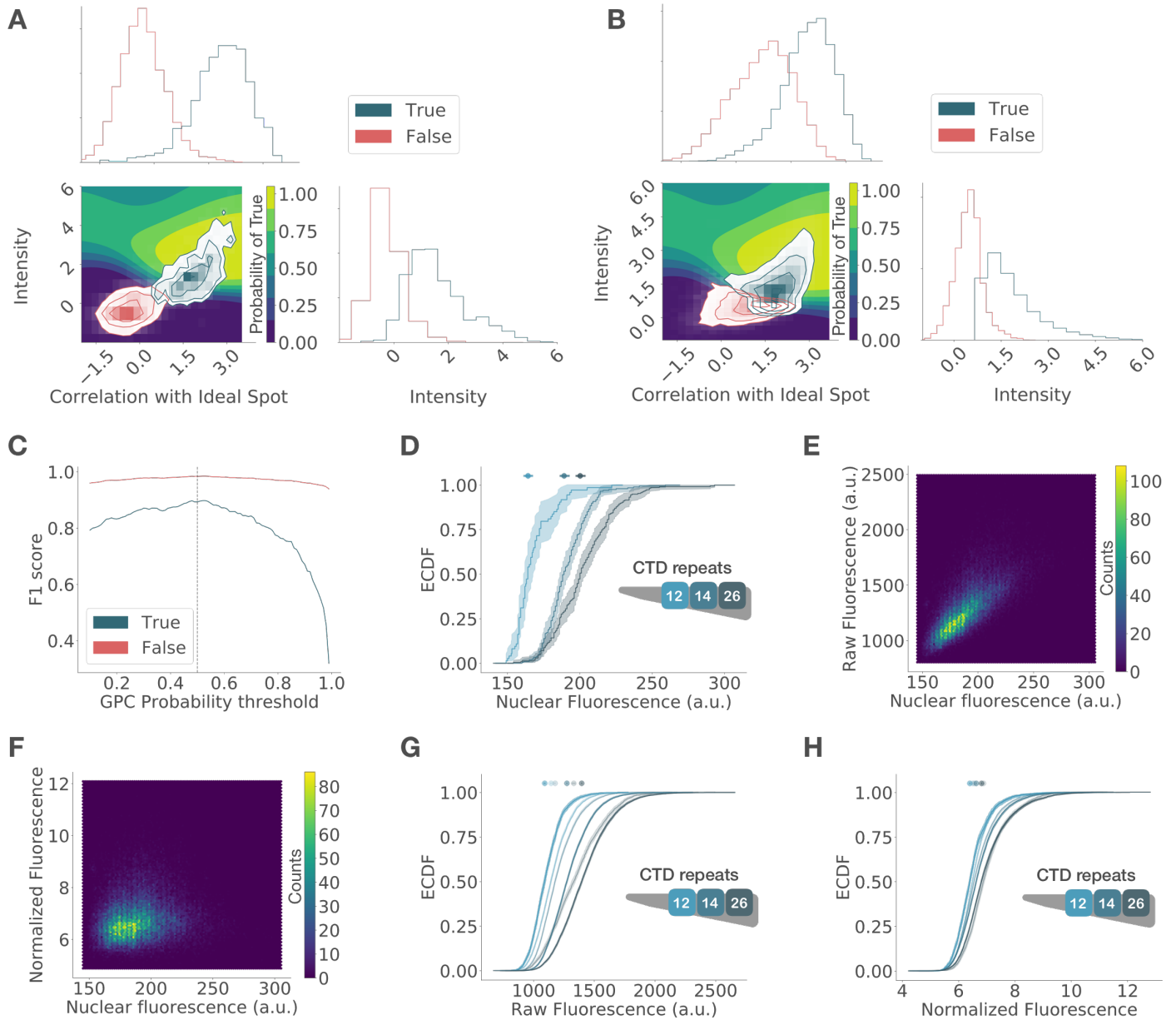
**RNA Pol II Length and Disorder Enable**

**Cooperative Scaling of Transcriptional Bursting**

**Porfirio Quintero-Cadena, Tineke L. Lenstra, and Paul W. Sternberg**

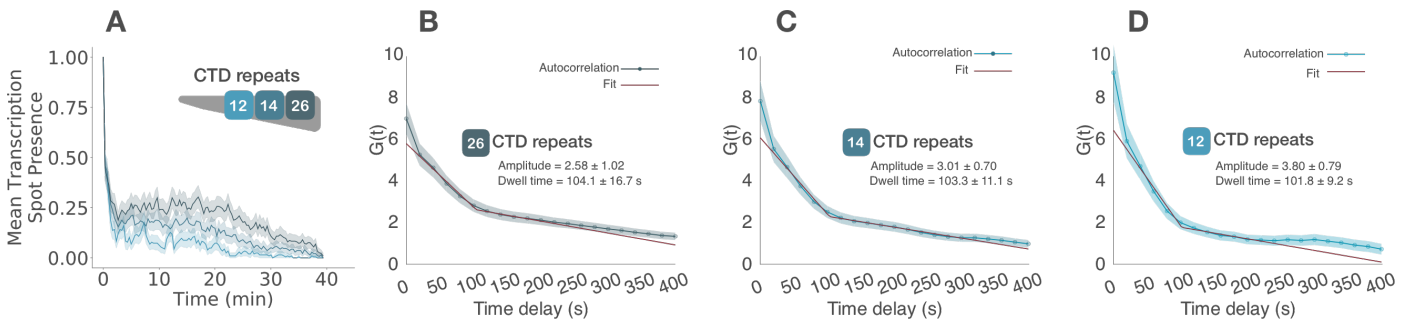


**Figure S1: CTDs share amino acid composition. Related to Figure 1.** CTDs were identified as the longest contiguous disordered region in RPB1 sequences. Only the longest protein per genus was considered. (A) Amino acid frequency sorted by mean abundance. Red dotted horizontal line indicates a uniform amino acid frequency of 1/20. Empirical cumulative distributions (ECDF) of net charges (B), aromaticity (C) and hydrophobicities (D) based on the grand average of hydropathy score (Kyte and Doolittle, 1982). Shaded area is bootstrapped 99% ECDF confidence interval (CI) and top markers show median with 99% CI.



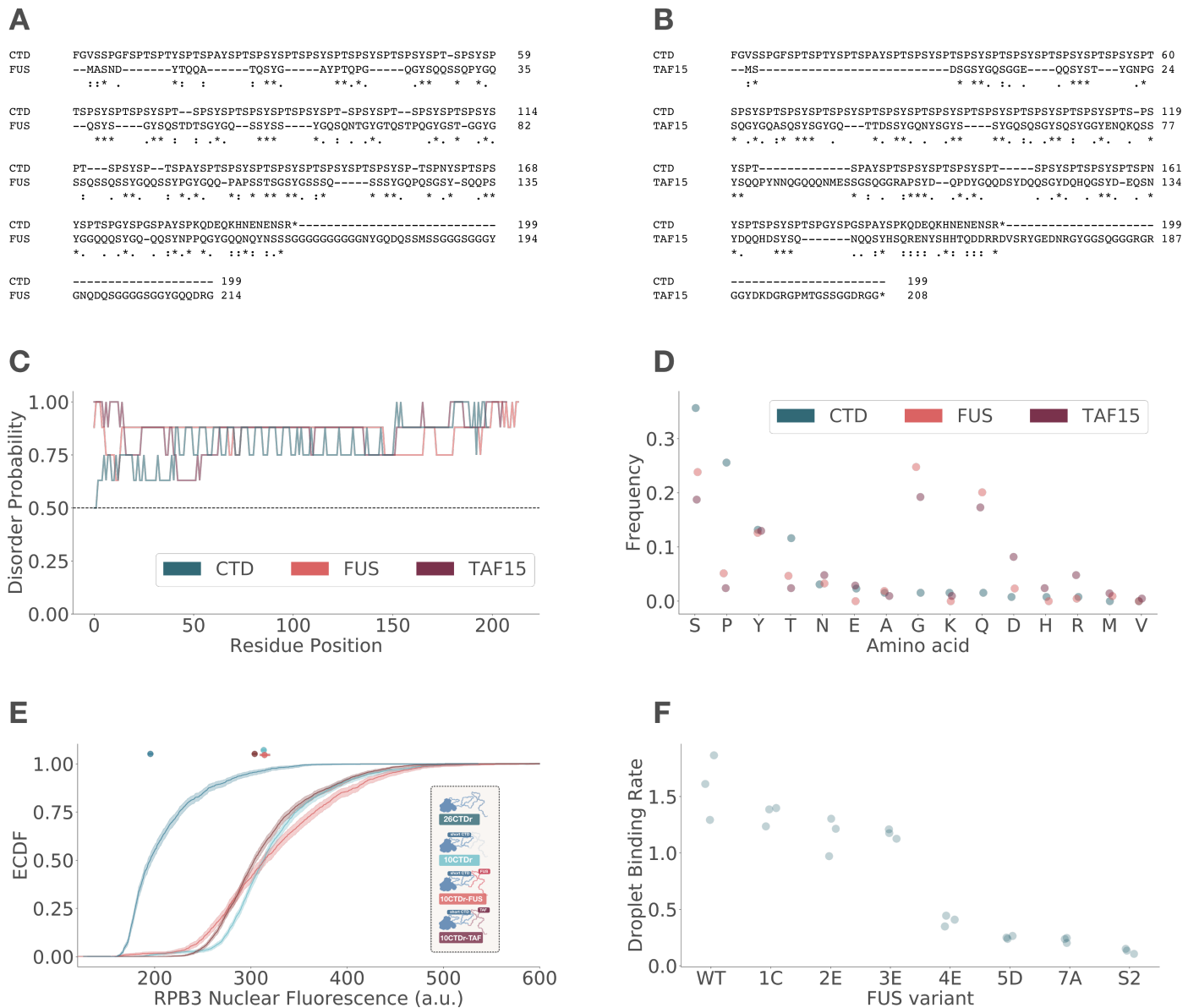
**Figure S2: Classification and normalization of PP7-GFP spots enables quantification of transcription dynamics and cross-strain comparisons. Related to figure 3.** Candidate spot images were obtained automatically using Trackpy's (Allan et al., 2018) peak detection algorithm. A sample of this image set was manually classified as True or False. For classification, spot images were represented using two features: correlation with an ideal spot (a single light point source blurred with a 2D Gaussian function) and intensity. (A) Histograms show the distribution of correlations (top) and intensities (right) of manually labeled spots. Left corner plot shows the joint distributions. This 2D data set was used to train a Gaussian-Process Classifier (GPC), resulting in the decision surface shown underneath, whose color indicates the probability of being a true spot. Candidate spots with a GPC probability above 0.5 were classified as True (B). This threshold was determined based on the change in the accuracy of classification (C), measured using the F1 score on a test set. The vertical dotted line indicates this probability threshold. Mutant strains show different PP7-GFP expression levels, as seen in the empirical cumulative distribution functions (ECDF) of mean nuclear fluorescence by strain (D). These differences result in a correlation observed in the hexagonal bin plot comparing mean nuclear fluorescence with raw spot fluorescence (E), which is removed after normalization (F). Normalized fluorescence is the ratio of spot fluorescence over mean nuclear fluorescence.

Figure S2: The efficacy of normalization can also be seen in the ECDFs of raw burst fluorescence by strain imaged with two laser intensities that artificially shift the intensity distributions of the same strains (G), which overlap after normalization (H). Transparency is used to indicate a different laser intensity. Shaded area is bootstrapped 99% ECDF confidence interval (CI) and top markers show median with 99% CI.

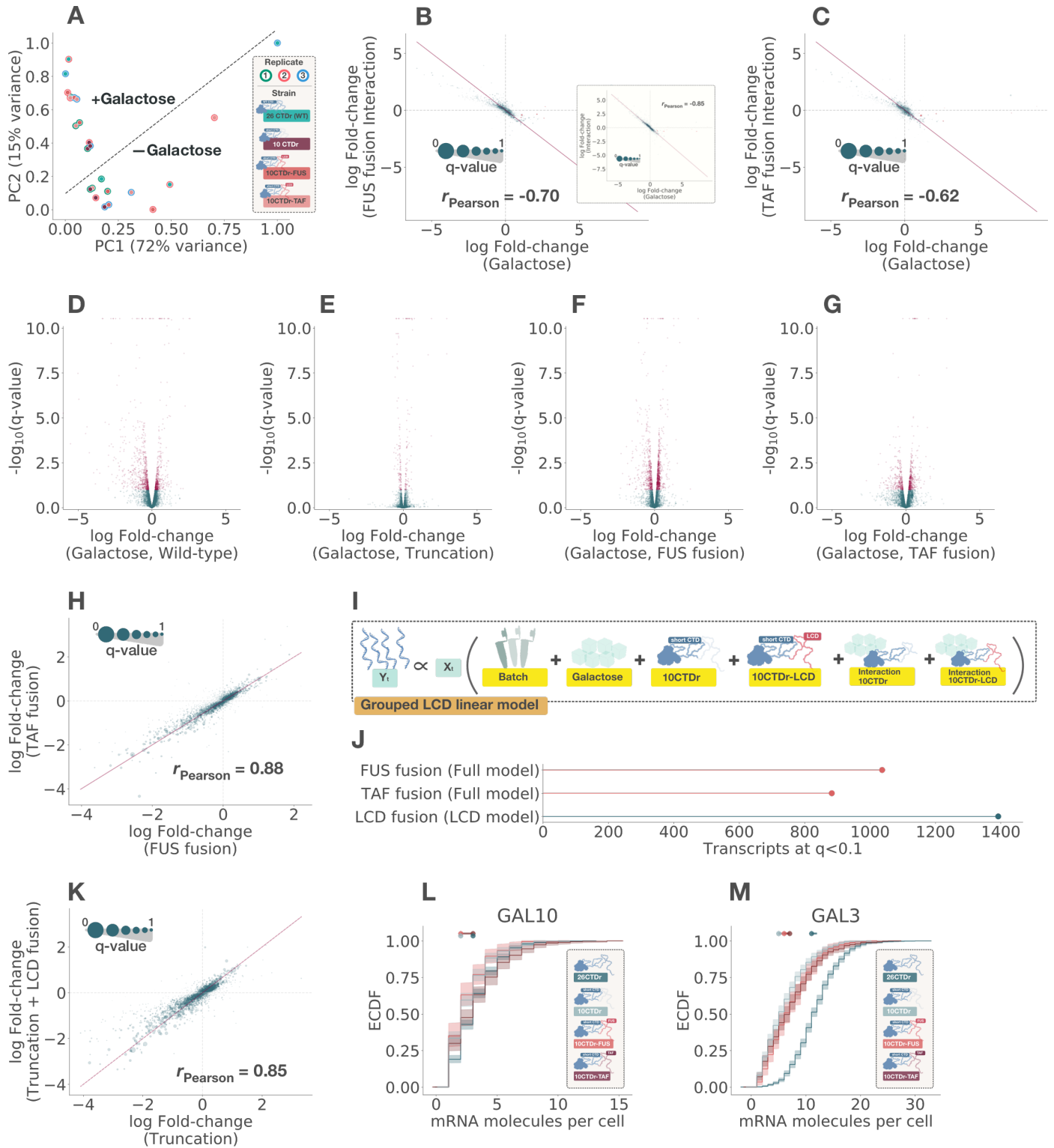


**Figure S3: Transcription burst frequency remains constant after activation and decreases with CTD truncation. Related to Figure 3.** (A) Mean aligned GAL10-PP7 boolean transcription traces. Boolean traces were obtained by marking with 1 and 0 the presence or absence of a transcription spot (TS), respectively. These traces were aligned and trimmed to begin with the first appearance of a TS and averaged over time, only considering cells that were active during the movie. These traces show the average frequency remains mostly constant over time and decreases with CTD length. Shaded area is bootstrapped 95% mean confidence interval. Frequency decay is also evident from an increase in amplitude, inversely related to frequency, in the autocorrelation of intensity traces corrected for non-steady-state effects in wild-type (B), 14 (C) and 12 (D) CTD $\Delta$  strains. Shaded area indicates standard error of the mean.





**Figure S4: FUS and TAF15 low complexity domains (LCD) are different in sequence but similar in amino acid composition to the CTD. Related to figures 4 and 5.** Protein alignments of FUS (A) and TAF15 (B) LCDs with yeast CTD. Disorder probabilities by residue (C) and amino acid frequency (D) in each of these proteins. Only amino acids present in at least one protein are shown. (E) Empirical cumulative distribution function (ECDF) of mScarlet-RPB3 in wild-type (26CTDr), truncated (10CTDr), and rescued strains (10CTDr fused to FUS or TAF15 LCD), from three biological replicates. Shaded area is bootstrapped 99% ECDF confidence interval (CI) and top markers show median with 99% CI. (F) *In vitro* droplet binding rates of FUS variants used in this study. These numbers are the slopes obtained from a linear regression of LCD-GFP binding to wild-type FUS LCD droplets, measured as droplet fluorescence intensity over time. Each point is an experimental replicate; data are from Kwon et al. (2013).



**Figure S5: Fusion of a CTD-truncated polymerase to FUS or TAF15 low complexity domains (LCD) results in convergent transcriptomes. Related to figures 2 and 4.** (A) Principal component analysis (PCA) with the first two PCs scaled to the range [0,1], which together explain 87% of the variance. Each strain has three biological replicates, indicated by edge color, and two conditions, separated by the dotted diagonal. Comparison of the log fold-change of each transcript induced by galactose and its interaction with 10CTDr fused with FUS (B) and TAF15 (C) LCDs. Inset shows comparison with 10CTDr interaction.

Figure S5: Volcano plots showing the estimated effect of galactose induction by transcript in wild-type (D), 10CTDr (E), FUS (F) and TAF15 (G) rescued strains. Transcripts that are below the significance threshold of 0.1 are shown in red. (H) Comparison of the log fold-change of each transcript resulting from FUS and TAF LCD fusion to 10CTDr truncated RNA Pol II under the full linear model shown in Figure 4B. The diagonal  $x = y$  is shown in red. Marker size of each point is inversely proportional to the q-value of the FUS coefficients ( $ms = -\log(q_{FUS})$ ); dotted gray lines reference no change at zero and the Pearson correlation is indicated. (I) Alternative linear model where FUS and TAF rescued strains are grouped together. This grouping results in a higher number of transcripts identified for LCD fusion under a q-value threshold of 0.1 than for individual coefficients (J). Using this model, (K) comparison of the log fold-change of each transcript resulting from truncation with and without LCD fusion. Empirical cumulative distribution function (ECDF) of mRNA molecules per cell from smFISH against GAL10 (L) and GAL3 (M) in wild-type, 10CTDr, FUS and TAF15 rescued strains. Shaded area indicates 99% ECDF confidence interval (CI); median with 99% CI is shown on top.

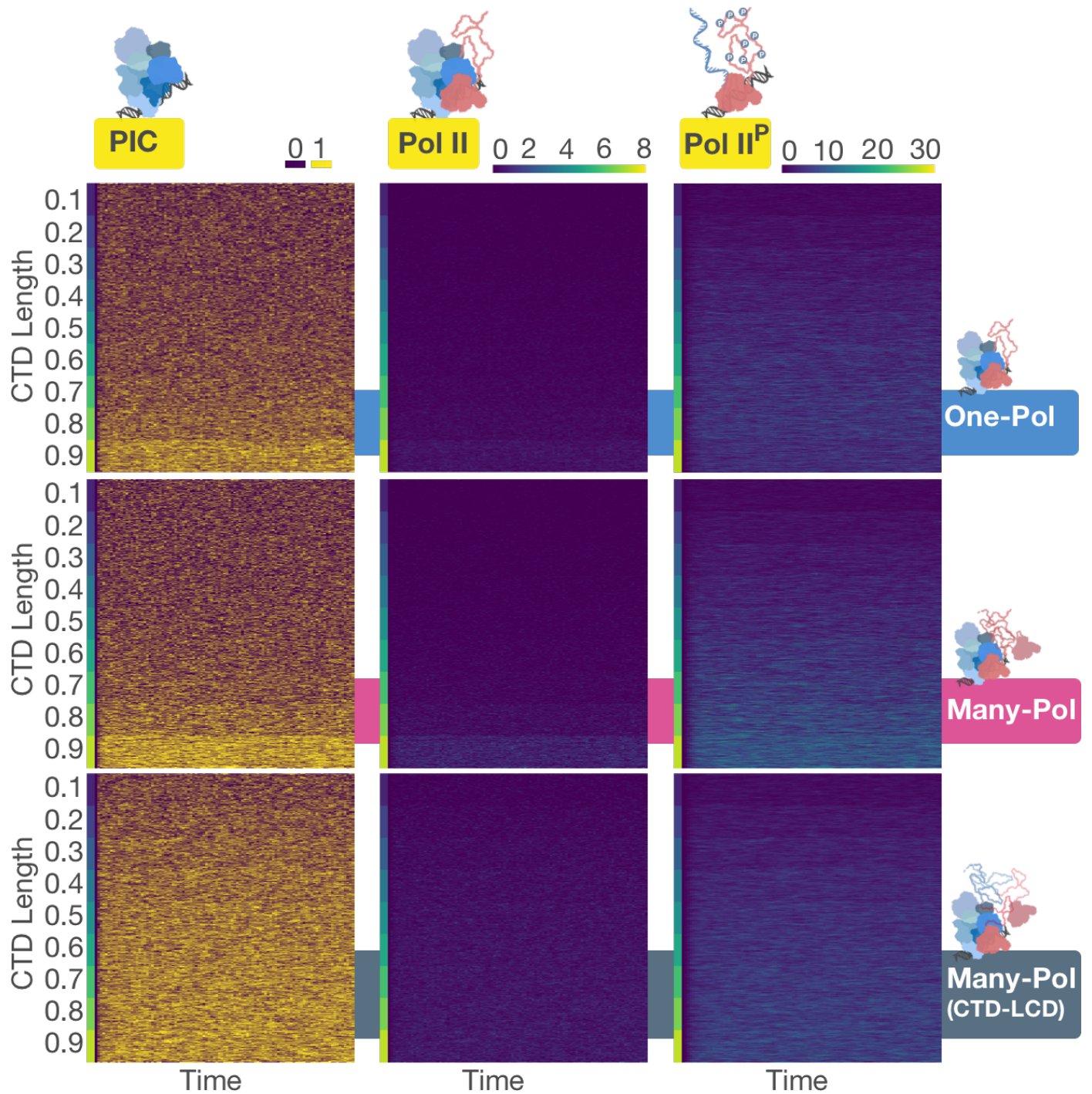


Figure S6: **Gillespie simulations yield traces akin to live transcription imaging. Related to figure 6.** Traces from stochastic simulations of PIC assembly states, number of PIC bound and phosphorylated (transcribing) polymerases for each model as a function of  $CTD_L$ , indicated with a colorbar to the left of each panel.

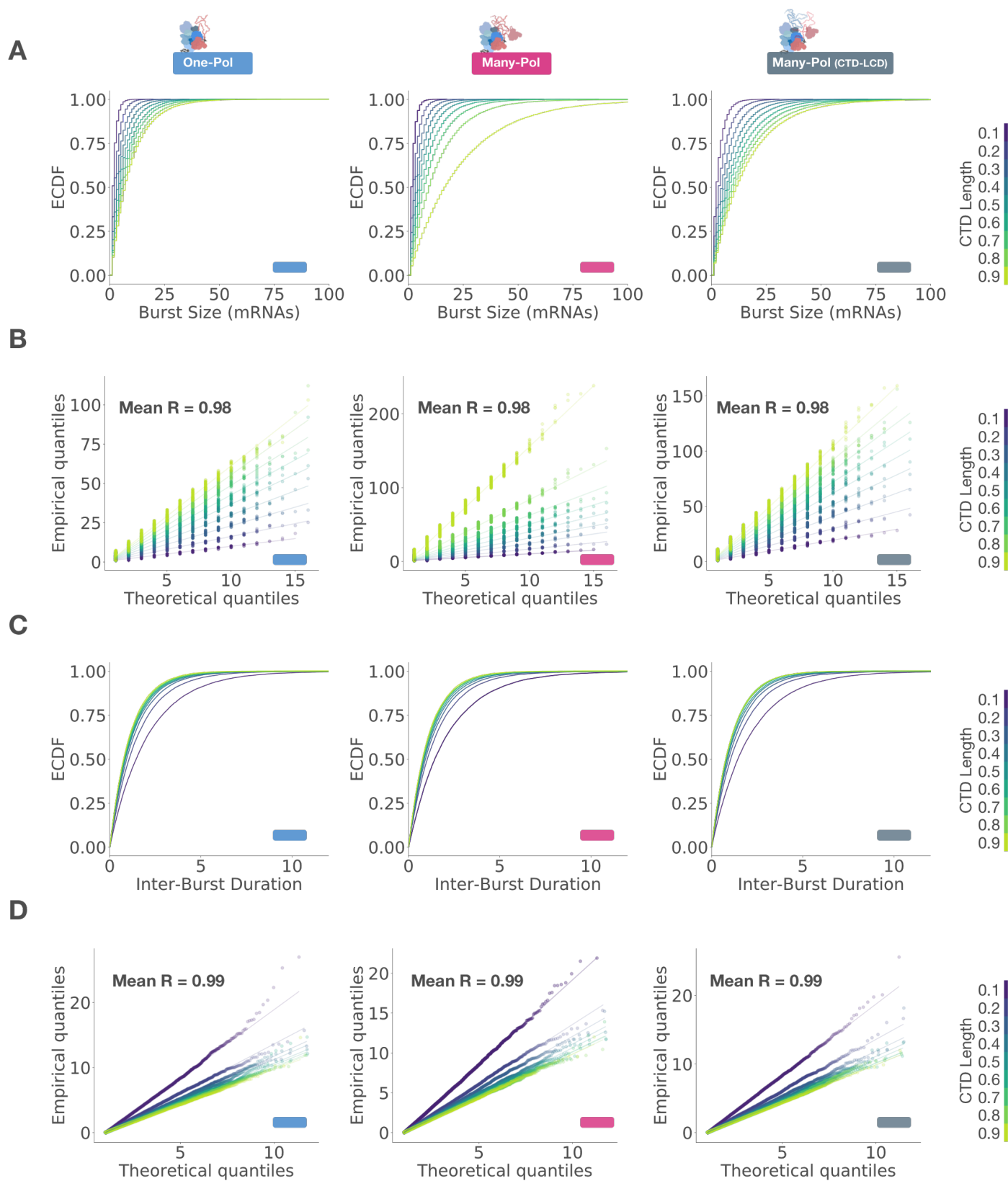
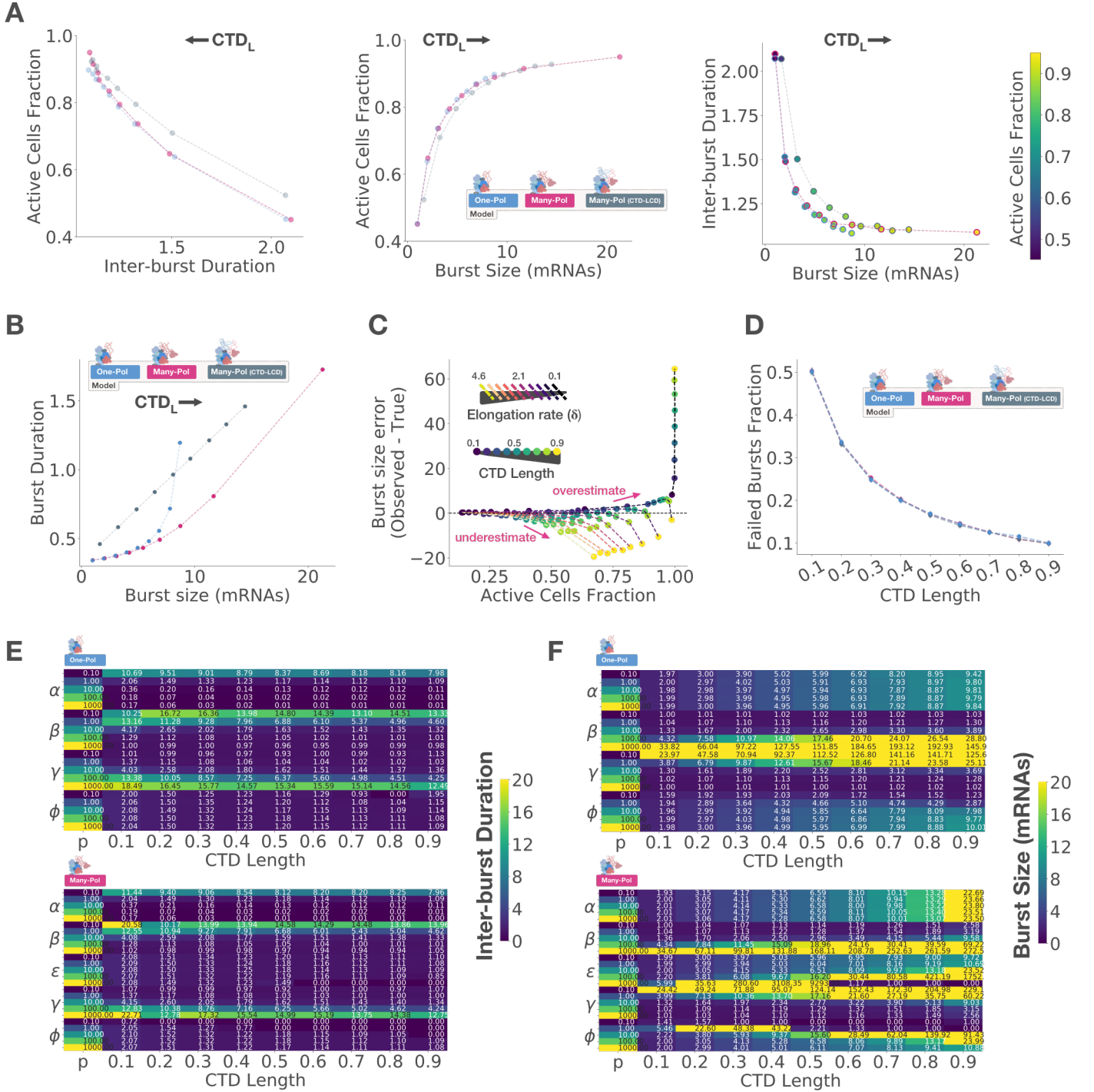


Figure S7:



**Figure S7: Transcription models produce geometric and exponential distributions of burst sizes and inter-burst durations, respectively. Related to figure 6.** (A) Empirical cumulative distribution functions (ECDF) of burst sizes by CTD length for each model. (B) Q-Q plots comparing quantiles from simulated distributions and a geometric distribution. Similarly, (C) ECDFs of inter-burst durations and (D) Q-Q plots comparing their quantiles with an exponential distribution. Each column comes from the model indicated by the color on top and the lower right corner in each plot. The mean of the square root of the coefficient of determination ( $R$ ) by model is indicated in each quantile comparison. CTD length is indicated by the color shown to the right of each row.



**Figure S8:**

**Figure S8: Parameter exploration with stochastic simulations provides insights into experimental observations. Related to figure 6.** (A) Comparison of the mean active cells fraction with means of inter-burst duration (left), burst size (middle) and both of these numbers (right) with increasing CTD length ( $CTD_L$ ) by model, indicated with color. Direction of CTD increase is indicated with an arrow on top of each plot. (B) Comparison of mean burst duration with mean burst size with increasing CTD length by model. (C) Comparison of the error in burst size estimate, computed as the difference between the means of the observed transcription site intensity and the true burst size, with the fraction of active cells as a function of  $CTD_L$  under the many-polymerases model. The elongation rate ( $\delta$ ) determines the time that a given mRNA spends bound to the transcription site and contributes to the observed intensity, thus influencing the fraction of active cells at a given time. A potential uncertainty in measuring burst size is that the decay in TS intensity could be a result of decreased burst frequency, given that GAL10 is transcribed in highly frequent bursts that could overlap in time and inflate the measured intensity of a burst. The trend in this analysis, coupled with the experimental range of active cells fractions (Figure 3E), suggests a scenario where the estimate from TS intensity lies between a slight overestimation to an underestimation of the true burst size; in the latter situation, inferring burst size from these data would be a conservative estimate. (D) Comparison of fraction of failed bursts, where an assembled preinitiation complex produced zero mRNAs before disassembly, as a function of  $CTD_L$  by model. Error bars indicate 99% bootstrapped confidence interval. Mean inter-burst duration (E) and burst size (F) as a function of  $CTD_L$  and individually varying parameter values, while the others are held constant, as indicated in the first left column of each heatmap. Colormap is artificially fixed to the range [0-20] for visualization purposes and actual numbers are shown in each cell. Model is indicated in the top left corner.